46 Pichersky, E. (1990) Nomad DNA – A model for movement and duplication of DNA sequences in plant genomes. *Plant Mol. Biol.* 15, 437–448
47 Stebbins, G.L. (1950) *Variation in the Evolution of Plants*, Columbia University Press
48 Sparrow, A.H. *et al.* (1976) Evoluion of genome size by DNA doublings. *Science* 192, 524–529
49 Werth, C.R. *et al.* (1991) A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *Am. Nat.* 137, 515–527
50 Bennett, M.D. *et al.* (1991) Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. London Ser. B* 334, 309–345
51 Hightower, R.C. *et al.* (1985) Divergence and differential expression of soybean actin genes. *EMBO J.* 4, 1–8
52 Baird, W.V. *et al.* (1987) A complex gene superfamily encodes actin in petunia. *EMBO J.* 6, 3223–3231
53 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
54 Zhang, L. *et al.* (1997) Gene expression profiles in normal and cancer cells. *Science* 276, 1268–1272
55 Fyrberg, E.A. *et al.* (1998) Functional nonequivalence of *Drosophila* actin isoforms. *Biochem. Genet.* 36, 271–287
56 Staiger, C.J. *et al.* (1997) Profilin and actin-depolymerizing factor: modulators of actin organization in plants. *Trends Plant Sci.* 2, 275–281
57 Gibbon, B.C. *et al.* (1997) Characterization of maize (*Zea mays*) pollen profilin function *in vitro* and in live cells. *Biochem. J.* 327, 909–915
58 Lopez, I. *et al.* (1996) Pollen specific expression of maize genes encoding actin depolymerizing factor-like proteins. *Proc. Natl. Acad. Sci. U. S. A.* 93, 7415–7420
59 Jiang, C.J. *et al.* (1997) The maize actin-depolymerizing factor, ZmADF3, redistributes to the growing tip of elongating root hairs and can be induced to translocate into the nucleus with actin. *Plant J.* 12, 1035–1043
60 Carpenter, J.L. *et al.* (1992) Preferential expression of an α-tubulin gene of *Arabidopsis* in pollen. *Plant Cell* 4, 557–571
61 Carpenter, J.L. *et al.* (1993) Semi-conservative expression of an *Arabidopsis thaliana* α-tubulin gene. *Plant Mol. Biol.* 21, 937–942
62 Kopczak, S.D. *et al.* (1992) The small genome of *Arabidopsis* contains at least six expressed α-tubulin genes. *Plant Cell* 4, 539–547
63 Snustad, D.P. *et al.* (1992) The small genome of *Arabidopsis thaliana* contains at least nine expressed β-tubulin genes. *Plant Cell* 4, 549–556

Reference added in press
64 Kandasamy, M.K., McKinney, E. and Meagher, R.B. The pollen specific actins in angiosperms. *Plant J.* (in press)

Making effective use of human genomic sequence data

Wonhee Jang
jang@ncbi.nlm.nih.gov

Hsiu-Chuan Chen
chenhc@ncbi.nlm.nih.gov

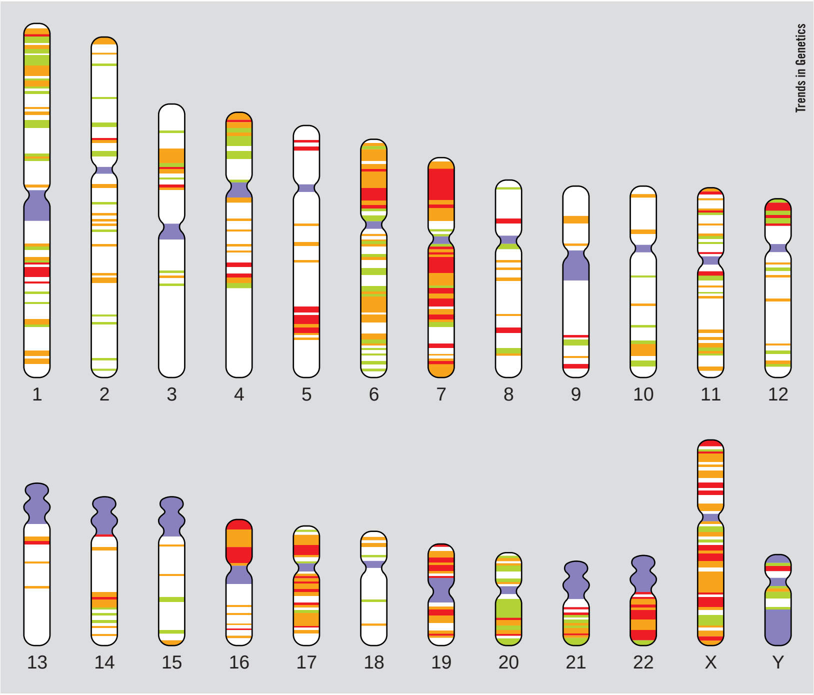
Hugues Sicotte
sicotte@ncbi.nlm.nih.gov

Gregory D. Schuler
schuler@ncbi.nlm.nih.gov

National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA.

The Human Genome Project has completed a successful three-year pilot project that has resulted in 280 875 kb of accurate, non-redundant human genomic sequence data being deposited in the public database (about 9% of the genome, shown graphically in Fig. 1). Recently, it has been announced that, as a first step towards accelerated completion of the genome, a 'working draft' sequence will be produced early in 2000 that will account for about 90% of the approximately three billion total bases[1]. This strategy might be likened to the authoring of a book – an initial working draft will be produced, followed by cycles of revisions in which details are added and errors corrected, ultimately leading to the final reference work. While it will be interesting to watch the developments in this fast-paced field over the next year or two, the practical benefit for most of us is the fact that huge amounts of DNA sequence will be pouring into the public databases. It is important for researchers to have early access to the genome because of its tremendous potential for accelerating biomedical research. We have developed an Internet resource (www.ncbi.nlm.nih.gov/genome/seq) to help make effective use of these data, by providing a simple means of browsing and searching of the data, together with background

FIGURE 1. Progress in human genome sequencing



Images of the 24 human chromosomes have been subdivided into segments of ~1 Mb in size and shaded to indicate available sequence data (at the end of March 1999). Additionally, completed regions are given different shades of orange and red to indicate the degree of continuity. Sequence data that are only of draft quality are shown in a contrasting light-green shade. Sequences were placed into segments based on sequence-tagged site (STS) content and combined radiation hybrid map data that has been converted to a uniform base-pair scale.

BOX 1. Genome sequencing centers

Baylor College of Medicine, USA
http://www.hgsc.bcm.tmc.edu

Center for Genetics in Medicine, USA
http://www.ibc.wustl.edu/cgm

Gesellschaft fur Biotechnologische Forschung mbH, Germany
http://www.gbf-braunschweig.de/welcomee.html

Genome Therapeutics Corporation
http://www.cric.com

The Institute for Genome Research
http://www.tigr.org

The Institute for Molecular Biology, Jena, Germany
http://www.genome.imb-jena.de

Max Planck Institute for Molecular Genetics, Germany
http://www.mpimg-berlin-dahlem.mpg.de

The Sanger Centre, Hinxton, UK
http://www.sanger.ac.uk

Stanford Human Genome Centre, USA
http://www.shgc.stanford.edu

University of Oklahoma Center for Genome Technology, USA
http://www.genome.ou.edu

University of Texas Southwestern Medical Center, USA
http://www.gestec.swmed.edu

University of Tokyo Human Genome Center, Japan
http://www.hgc.ims.u-tokyo.ac.jp

University of Washington Genome Center, USA
http://www.genome.washington.edu/UWGC

University of Washington Multimegabase Sequencing Center, USA
http://www.chroma.mbt.washington.edu/msg_www

US Dept of Energy Joint Genome Institute
http://www.jgi.doe.gov

Washington University Genome Sequencing Center, USA
http://www.genome.wustl.edu/gsc

Whitehead Institute for Biomedical Research, USA
http://www.genome.wi.mit.edu

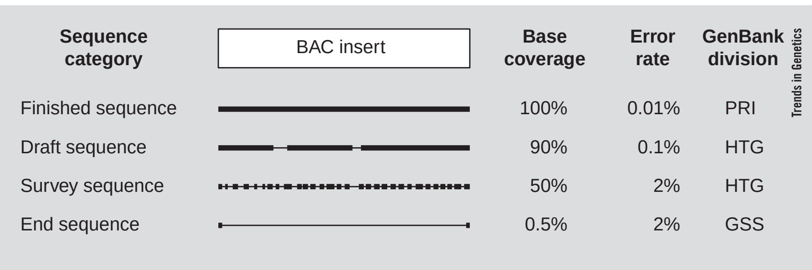
(This list includes only centers that have contributed at least 1000 kb of finished sequence. Several of these web sites have been recently reviewed by Pruitt[8].)

information and an assessment of overall sequencing progress.

Unlike a real book, which would never be published while still in draft form, the genomic sequence – the 'Book of Life' – is made available to the public, even as the pages are written. The international sequencing community has set for itself a policy of immediate data release, in which shotgun sequence assemblies at least 2 kb in length are made publicly available within 24 hours. These sequences are deposited in the GenBank, EMBL and DDBJ public sequence databases and are then redistributed to the world on a daily basis. Data are also made available on individual genome center web sites, usually augmented with mapping data and information on work in progress (Box 1). Increasingly, genomic sequence entries represent individual bacterial artificial chromosome (BAC) clones, those being the current sequencing target of choice. But sequences of BACs can take on a variety of forms that, owing to their unique properties and uses, are segregated into different GenBank divisions (Fig. 2). For example, having finished sequences in the primates (PRI) division makes it easier for users to distinguish them clearly from other sequences that have lower accuracy and that might contain gaps. The properties of a draft sequence might change over time, as more experience is gained and production strategies evolve. Although several important factors must be balanced, it is possible to obtain high-quality data for over 90% of the bases of a clone insert[2].

A new category of genomic sequence entry is the low-pass survey sequence. Such a sequence consists of a relatively small number of single-pass sequencing reads whose order is unknown. Survey sequences are used by the genome centers for the dual purposes of evaluating the quality of shotgun libraries and verifying that a clone is not excessively redundant with previously sequenced regions. For users, the utility of this sort of sequence is that, starting with a cDNA sequence of interest, it is possible to perform a database search and find a BAC clone that contains the gene. The final category is BAC-end sequences,

FIGURE 2. Categories of BAC sequences found in GenBank



In each case, a typical pattern of coverage is illustrated, together with an approximate percentage of the bases covered. For the draft sequence, the coverage and error rate can differ from that shown, depending on the exact strategy that is ultimately adopted. Coverage for the survey sequence will vary substantially, depending on the size of the insert and whether one or two plates of subclones is sequenced. Error rates given for survey and end sequences are based on those typically seen for single-pass sequences. Finished sequences are stored in the GenBank division that is appropriate for the given organism, which in the case of *Homo sapiens* is primates (PRI). Finished *Mus musculus* entries would be in rodents (ROD), *Drosophila melanogaster* would be in invertebrates (INV), and so forth. Draft and survey sequences are stored in the high-throughput genomic (HTG) division, which is used for sequences that are still undergoing revision. Upon completion, sequences are reassigned to PRI and the same accession number is kept throughout the revision process. Bacterial artificial chromosome- (BAC-) end sequences are stored in the genome survey sequence (GSS) division. For more information on GenBank divisions, see Ref. 9.

TABLE 1. Distribution of contig sizes

Size range (kb)	Contigs	Aggregate size (kb)	Proportion of total (%)
<30	47	676	0.2
30–100	453	28 418	10.1
100–250	915	141 295	50.0
250–500	182	59 350	21.1
500–1000	46	31 741	11.3
>1000	15	19 394	6.9
Total	1658	280 875	100.0

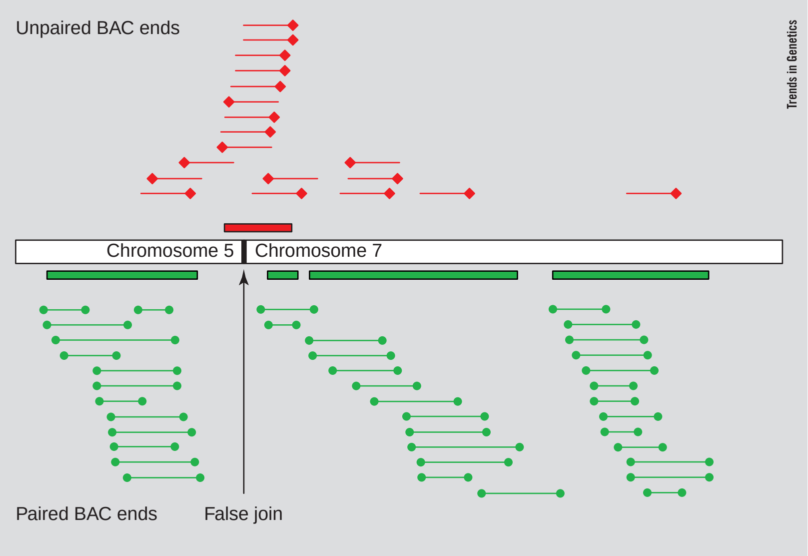
Contig size data were collected on 31 March 1999. For the latest information, see www.ncbi.nlm.nih.gov/genome/seq. Constructed sequences are given accession numbers from a new series in which identifiers begin with the letters NT (e.g. NT_000001). All entries are freely available through the web interface.

which are being produced in large numbers through a project funded by the US Department of Energy. These pairs of short (~400 bases) sequences, which are separated by a defined distance (the size of the BAC insert), have a number of uses, including the identification of clones that would extend or bridge existing islands of sequence[3].

The reconstruction of the complete human genome will involve merging the individual clones into longer blocks of contiguous sequence. At the National Center for Biotechnology Information (NCBI), we have already begun this task, drawing primarily upon clone-overlap information provided by the genome centers or given in the sequence annotation. This has been augmented by computational analysis, making use of the electronic PCR program[4] to find mapped sequence-tagged site (STS) markers in the sequences and BLAST (Ref. 5) to identify probable overlaps. All expected overlaps are confirmed by sequence alignment before constructing the final sequence entries. As shown by the size distribution of the segments (Table 1), several long sequences are already available, including 15 that exceed 1000 kb in length. Overall, about 90% of the total bases fall into segments that are at least 100 kb. All of the sequences are available for browsing and downloading through the web interface.

Quality assurance will be critical for the final reconstructed genome and there are many advantages to monitoring quality constantly as the Genome Project progresses[6]. Discussions of sequence quality often center around the accuracy of determining each individual base, but quality can be viewed at other levels as well. For example, restriction-site analysis has been used to verify that the sequence obtained for an individual clone is consistent with its physical DNA (Ref. 7). But additional checking is needed to ensure that the reconstructed sequence is colinear with the genome. Discontinuities can be caused by joining clone sequences together in an incorrect order or by actual rearrangements of cloned DNA. We have developed methods for using BAC-end sequences to check for reconstruction errors in genomic sequences (Fig. 3). Briefly, it is expected that paired BAC end sequences should align to the genome with a spacing that is consistent with the size of the clone insert. Regions are considered to be confirmed if they are spanned by at least two BACs with appropriately paired ends. A prevalence of unpaired end sequences is indicative of discontinuities. To date, very few problems have been found in real data and they have been resolved by consultation with genome centers. To illustrate the technique, a hypothetical error case was prepared by intentionally joining sequences from different regions of the genome. As shown in Fig. 3, the point of the incorrect join is revealed by a large number of unpaired BAC ends corresponding to a gap in the coverage by paired ends.

FIGURE 3. Validating genomic contig sequence



Two different contigs (NT_000169, 476 kb, and NT000431, 1040 kb) were intentionally joined together. The false assembly junction at 476 kb can be identified by a paucity of BAC with both ends aligning and a marked increase in the number of unpaired BAC ends. Other regions of the contig were covered 1–10-fold with BACs having paired ends. Green boxes indicate regions covered by at least two paired BAC ends. Red box indicates region with at least four unpaired BAC ends.

We have briefly discussed several issues relating to public genomic sequence entries, their use in constructing larger constructs, and methods for verifying their correctness. Additional details are available on the Human Genome Sequencing web site. It is possible to find regions of interest by using simple text queries and downloading any sequences found. A graphical viewer shows how the sequences were constructed and indicates regions that have been confirmed by paired BAC ends. Users can perform BLAST searches against the human genomic sequence and, in addition to traditional sequence-alignment output, they can obtain a graphical view of the positions of matching regions relative to other features in the larger genomic segment. For example, using a mouse cDNA sequence as a query, a putative human homolog can be identified in a region that also contains polymorphic STS markers, which might be useful tools for further genetic analysis. Although the final reference human genomic sequence will be several more years in the making, there is no need to delay in making effective use of the data that exist today.

References

1 Pennisi, E. (1999) Academic sequencers challenge Celera in sprint to the finish. *Science* 283, 1822–1823
2 Bouck, J. *et al.* (1998) Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* 8, 1074–1084
3 Siegel, A. *et al.* (1999) Analysis of sequence-tagged-connector strategies for DNA sequencing. *Genome Res.* 9, 297–307
4 Schuler, G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.* 7, 541–550
5 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
6 Olson, M. and Green, P. (1998) A 'quality-first' credo for the Human Genome Project. *Genome Res.* 8, 414–415
7 Wong, G.K. *et al.* (1997) Multiple-complete-digest restriction fragment mapping: generating sequence-ready maps for large-scale DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5225–5230
8 Pruitt, K.D. (1998) WebWise: web sites of the Human Genome Project. *Genome Res.* 8, 1109–1111
9 Ouellette, B.F. and Boguski, M.S. (1997) Database divisions and homology search files: a guide for the perplexed. *Genome Res.* 7, 952–925